IET Computer Vision

Special Section: Computer Vision in Healthcare and Assisted Living

IET Journals
The Institution of Engineering and Technology

Energy expenditure estimation using visual and inertial sensors

ISSN 1751-9632
Received on 15th February 2017
Revised 19th September 2017
Accepted on 23rd September 2017
E-First on 27th October 2017
doi: 10.1049/iet-cvi.2017.0112
www.ietdl.org

Lili Tao[1,2], Tilo Burghardt[1], Majid Mirmehdi[1] ✉, Dima Damen[1], Ashley Cooper[1], Massimo Camplani[1], Sion Hannuna[1], Adeline Paiement[1], Ian Craddock[1]

[1]*SPHERE, Faculty of Engineering, University of Bristol, Bristol, UK*
[2]*Centre for Machine Vision, Bristol Robotics Laboratory, University of the West of England, Bristol, UK*
✉ *E-mail: M.Mirmehdi@bristol.ac.uk*

Abstract: Deriving a person's energy expenditure accurately forms the foundation for tracking physical activity levels across many health and lifestyle monitoring tasks. In this study, the authors present a method for estimating calorific expenditure from combined visual and accelerometer sensors by way of an RGB-Depth camera and a wearable inertial sensor. The proposed individual-independent framework fuses information from both modalities which leads to improved estimates beyond the accuracy of single modality and manual metabolic equivalents of task (MET) lookup table based methods. For evaluation, the authors introduce a new dataset called *SPHERE_RGBD + Inertial_calorie*, for which visual and inertial data are simultaneously obtained with indirect calorimetry ground truth measurements based on gas exchange. Experiments show that the fusion of visual and inertial data reduces the estimation error by 8 and 18% compared with the use of visual only and inertial sensor only, respectively, and by 33% compared with a MET-based approach. The authors conclude from their results that the proposed approach is suitable for home monitoring in a controlled environment.

1 Introduction

The term ‘energy expenditure’ refers to a human's calorific uptake over time, which is one commonly used single metric to quantify physical activity levels. It is an important determinant in understanding the development of chronic diseases, such as obesity and diabetes. Current evidence-based guidelines [1] indicate that people who are regularly physically active have a 20–40% lower risk of developing conditions such as cardiovascular disease and type 2 diabetes than those who are inactive, and suggest that adults should accumulate at least 150 min of moderate intensity physical activity each week or 75 min of vigorous activity, or a combination of the two. Most research into estimating and understanding calorific expenditure focuses on coarse energy totals over longer time segments or relates to specific activities only, such as walking and running, which generally occur outside the home.

Yet, very little attention has been paid on how activities of normal daily living in an indoor environment can be quantified and understood in terms of energy expenditure. Traditionally, physical activity levels have been measured in Metabolic Equivalents of Task (MET) [2], where a fixed value is assigned to each activity, e.g. 1 MET corresponds to energy expended at rest. However, the method is highly unreliable due to the fact that the activities are monitored using self-report approaches, such as questionnaires and occasional clinical check-ups.

There are various approaches that reliably estimate human energy expenditure via respiratory gas analysis, including both direct and indirect methods. Direct calorimetry measures, such as a sealed respiratory chamber [3], produce accurate outputs, but require a laboratory-based environment. Indirect calorimetry, on the other hand, measures energy expenditure based on inspired and expired respiratory gas flows, volumes and concentrations of oxygen consumption and carbon dioxide production. Some of these measurement devices are portable, less invasive and can produce accurate readings. They form the measurement standard for non-stationary scenarios where the person can move freely. Nevertheless, participants in experiments are required to carry gas sensors and wear a breathing mask [4].

Recently, with an increasing number of wearable devices becoming available, approximating the energy expenditure using inertial sensors has become a popular monitoring choice due to its low cost, low energy consumption, and data simplicity. Acceleration reflects a relation between motion and energy expenditure, thus tri-axial accelerometers are the most broadly used inertial sensors [5]. Recent studies show that more sensors could be involved in the task: breath rate, chest and arm skin temperature also show the correlation with energy expenditure via estimating the oxygen consumption [6]. The data could be obtained by a heart rate monitor and thermometers.

Vision-based systems, as alternative approximative sensors, do not require the wearing of extra devices. In fact, they are already a key part of home entertainment systems today [7], where RGB-Depth sensors allow for a rich and fine-grained analysis of human activity for purposes such as gaming within the field of view. Recent advances in computer vision have now opened up the possibility of integrating these devices seamlessly into home monitoring and assisted living systems [8–10].

Simultaneous utilisations of visual and inertial sensors are not common today, but are receiving growing attention in various areas, including action recognition [5], gesture recognition [11], robotics [12], augmented/virtual reality [13], and assistive technologies applications, such as fall detection [14], food preparation [15] and in a general ambient assisted living system [16]. Although employing multi-modal sensors has the advantage of complementing shortcomings of individual modalities, wearing a multitude of sensors can cause user acceptance issues.

With this in mind, in this paper we propose a framework for estimating energy expenditure in living environments based on a non-intrusive RGB-Depth visual sensor and two inertial sensors – worn on the wrist and waist – backed up in experiments by simultaneously taken indirect calorimetry measurements based on the measurement of oxygen consumption and carbon dioxide production for an accurate ground truth provision. This is a new application and to the best of our knowledge no dataset of a similar setup with reliable and accurate ground truth exists. Thus, in order to quantify the performance of the proposed system, we present a new dataset, *SPHERE_RGBD + Inertial_calorie*, for calorific expenditure estimation collected within a home environment. The dataset contains 11 common household activities performed over

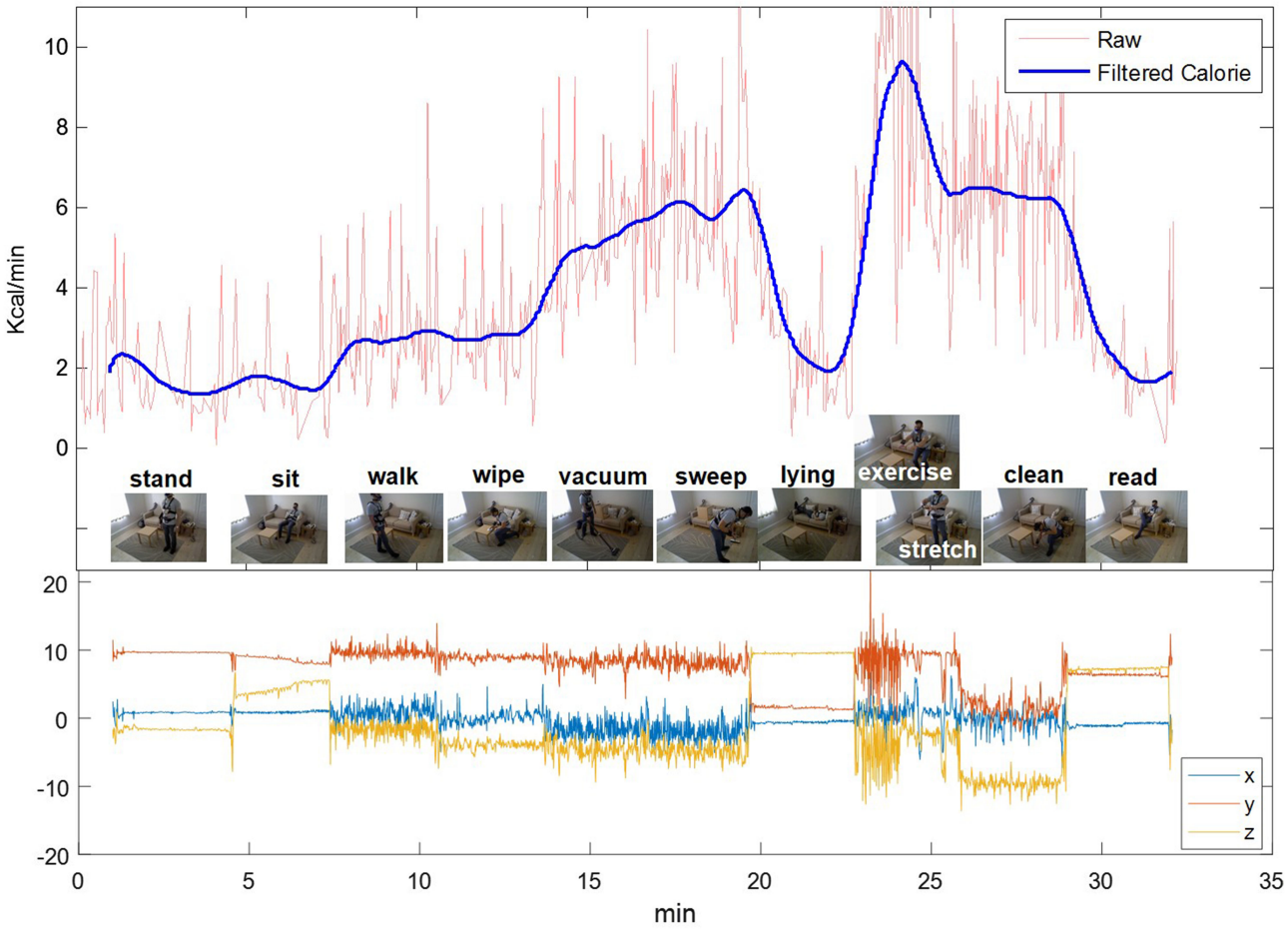


Fig. 1 *Ground truth example sequence. Top: raw per breath data (red) and smoothed COSMED-K4b2 calorimeter readings (blue) and sample colour images corresponding to the activities performed by the subject. Bottom: three-axis acceleration signals from the waist-wear sensor*

up to 20 sessions, lasting up to 30 min for each session, in each of which the activities are performed continuously. The experimental setup consists of an RGB-Depth Asus Xtion camera mounted at the corner of a living room, two accelerometer sensors, and a COSMED K4b2 [4] indirect calorimeter for ground truth measurement (see Fig. 1). The *SPHERE_RGBD + Inertial_calorie* dataset is publicly released (The dataset is available online at http://doi.org/cc5k.).

This paper is built on our recent work in [17–19], with significant extensions and improvements. Tao *et al.* [17] introduced a fusion framework for recognising human daily activity using visual and inertial sensors. The work did not address the issue of energy expenditure estimation. Tao *et al.* [19] proposed a framework for calorific expenditure estimation using only a visual sensor, thus, there was no sensor fusion involved. In [18], we presented a system which allows real-time prediction for activity intensity levels, relying on light-weight bounding box features. This makes the method unable to produce precise calorific expenditure values. In this work, we have improved the feature representation for both inertial and visual sensor data by considering spatial and temporal information at the same time, and investigated both early and late fusion approaches of the data from these sensors. The key contributions of this work are as follows. (i) We propose a first-ever framework for the estimation of calorific expenditure from a RGB-Depth sensor and inertial wearable sensors. There is no work published on visual-inertial energy expenditure estimation, there being only very few works that offer purely vision-based estimation [18, 20, 21]. (ii) We improve the feature representation for both inertial and visual features in the previous fusion framework in [17] by extracting rich, multi-level information to give improved estimation accuracy. (iii) We introduce a new dataset, linking more than 10 h of RGB-Depth video data and inertial sensor data to ground truth calorie readings from indirect calorimetry based on gas exchange. (iv) We present a comparative study on the utility of both visual and inertial data when estimating energy expenditure in a living environment. The visual sensor and inertial sensors are evaluated individually first, followed by an evaluation of two fusion approaches. The rest of the paper is organised as follows. Section 2 presents the background and work related to our study. Section 3 describes the proposed framework for estimating energy expenditure from RGB-Depth and inertial sensors alone, as well as in fusion. The experimental setup and the results are presented in Section 4, followed by a discussion and our conclusions in Section 5.

2 Related work

2.1 Inertial sensors

Acceleration, angular velocity, and rotation signals from wearable devices have been used for human action recognition [22], and are popular in healthcare-oriented applications, such as in fall detection systems [23] and medication adherence monitoring systems [24]. Inertial sensors can offer particularly low-cost and ubiquitous monitoring solutions for physical activities. Techniques that can control computational complexity, power consumption, and improve the unobtrusiveness of the wearable computers [25] are applicable to many systems including the one at hand. Here, we first discuss inertial sensor feature extraction methods described in the literature, followed by an outline of existing models of energy expenditure estimation based on them.

Feature representation: Different features extracted from inertial sensor devices have been considered ranging from raw signal samples to high-level descriptors. Raw time series data from accelerometers is most often provided as triples of scalars, where each scalar corresponds to acceleration in one of three orthogonal spatial dimensions. The same fundamental structure applies to angular velocity signals and orientation signals of three directions. There is no computational burden associated with feature extraction when the raw data is used.

However, raw data may not expose enough discriminative structure to achieve high performance on specific classification tasks. Instead, statistical features may be extracted from each of the three axes, where sensor signal sequences are often partitioned into temporal windows over which features are generated. All features extracted from a temporal window are then concatenated to form a single combined descriptor vector.

Commonly used features include the first- and second-order statistics, namely the mean and variance [26]. In [17], apart from these commonly used features, correlation measures between each axes pair are also extracted. Basic statistical measures are computationally efficient and are able to capture structural patterns in inertial data. The feature descriptor can be further quantised into a number of codewords, such as in [27]. Approaches based on deep learning are currently being explored to create more generalised learning methods that generate features directly from the input data and promise to optimise performance further [28].

Energy expenditure estimation: The first automatic methods for inertial energy expenditure estimation [29] were count-based estimation systems applied by fitting a single regression model to all the data regardless of what activity was being performed. However, systems that map from a single wearable to calorie values struggle to accurately estimate the intensity of physical activity across a range of actions. For example, some actions involving only upper or lower body movements are difficult to be recognised via a single wearable device, and therefore a high estimation error [30] occurs. Different activities may require different models to represent them. Activity-specific (AS) methods split the estimation process into two steps, where activity groups are detected and classified first, and only then an AS model is applied to estimate energy expenditure. MET lookup tables are the most common approach to perform the latter, where a static MET value is assigned from a compendium on physical activities [2] to each one of the clusters of activities [31]. However, METs-based approaches neglect any transitional effects of activities (continued calorie expenditure after rigorous activity has finished), and they overlook the fact that even the activities in same cluster can be performed at varying intensities, for example, walking at different speeds, or body exercise with different intensity.

An attempt to model the transition between activities was proposed in [32], where an accelerometer and a heart rate sensor were used and the transition between sedentary, household activities, and walking were modelled. The work in [33] shows that by using data from multiple inertial sensors one can more accurately predict energy expenditure, although the limitations of wearable devices are considerable particularly with respect to accuracy as emphasised in [17].

Accelerometer feature descriptors are often formed within a temporal window. This brings out another concern that the window sizes are set usually at <10 s in existing works [6, 32]. The length of window would significantly affect the results. It should be short enough to recognise activities as local temporal information are more descriptive, but long enough to predict calorie values since current energy expenditure strongly depends on previous activity intensity level.

2.2 Visual sensors

Visual sensor based techniques have emerged over recent years for which there exists a significant body of literature describing the inference of activities from two-dimensional (2D) colour intensity imagery [34]. Meanwhile, the increasing availability of depth-measuring sensors, especially the introduction of the Microsoft Kinect, has generated an opportunity for utilising depth in conjunction with traditional RGB camera data allowing for richer and more fine-grained analysis of human activity [7]. Applying computer vision techniques to help with the diagnosis and management of health and wellbeing conditions has gained significant momentum over the last years [35]. However, studies on energy expenditure using visual sensors have been relatively limited. Our work explores this field further and builds on several relevant subject areas in computer vision.

Visual feature representation: The visual trace of human activity in video forms a spatio-temporal pattern. To extract relevant properties from this for the task at hand, one aims at compactly capturing this pattern and highlighting important aspects related to the properties of interest. Assuming that both body configuration and body motion [36] are relevant to infer calorific uptake, the pool of potential features is large – ranging from local interest point configurations [37], over holistic approaches like histograms of oriented gradients and histograms of motion information [17], to convolutional neural network features [38].

Motion information in the first place could also be recovered in various ways, e.g. from RGB data using optical flow *or* from depth data using 4D surface normals [39]. Whilst a composition of these features via concatenation of per-frame descriptors is straight forward, this approach suffers from the curse of dimensionality and unaffordable computational cost. Sliding window methods [40], on the other hand, can limit this by predicting current values only from nearby data within a temporal window. Further compaction may be achieved by converting large feature arrays into a single, smaller vector with a more tractable dimension count via, for instance, bags of visual words [41], Fisher vectors [42], time series pooling [43], or the features extracted from convolutional neural networks [44]. In summary, the challenge of feature representation will require capturing visual aspects relevant to calorific expenditure, whilst limiting the dimensionality of the descriptor.

Activity recognition: There exists a significant body of literature describing the inference of activities from 2D colour intensity imagery [34], RGB-Depth data [7], and skeleton-based data [45]. Knowledge about the type of activity undertaken has been shown to correlate with the calorific expenditure incurred [2]. In alignment with Fig. 2*a*, we will argue in this work that an explicit activity recognition step in the vision pipeline can, as an intermediate component, aid the visual estimate of energy uptake.

Energy expenditure estimation: 2D video has recently been used by Edgcomb and Vahid [20] coarsely to estimate daily energy expenditure. In their work, subjects are first segmented from the scene background. Changes in height and width of the subject's motion bounding box, together with vertical and horizontal velocities and accelerations, are then used to estimate calorific uptake. Tsou and Wu [21] take this idea further and estimate calorie consumption using full 3D joint movements tracked as skeleton models by a Microsoft Kinect. We note, however, that both of the above methods use wearable accelerometry as the target ground truth, which in fact does not provide an accurate benchmark; skeleton data is commonly noisy and currently only operates reliably when the subject is facing the camera. This limits applicability in more complex, in-the-wild, visual settings as, for instance, contained in the *SPHERE_RGBD + Inertial_calorie* dataset. Our recent work in [19] introduced a visual based framework for estimating calorific expenditure in a home environment, and we then extended it to be able to estimate physical activity intensity levels in real time [18]. Although the method is practically applicable to more complex settings, the light-weight features extracted from bounding boxes (velocity vector and the ratio of height and width of the bounding box) can only help make a gross estimate of calorific expenditure. In this paper, instead of using only simple bounding box features, we simultaneously collect RGB and depth imagery and then encode appearance and motion features via spatial pyramids. The temporal information is encoded using a pyramidal temporal pooling with multiple pooling operators. This has the aim of extracting rich, multi-level information to give improved estimation accuracy, whilst maintaining applicability to detect complex human activities.

2.3 Sensor fusion

It is reasonable to expect that the use of multiple sensor types would improve the overall performance compared with single sensor settings, since sensors may complement the limitations of each other. Given an accurate temporal synchronisation between the different modality sensors, learning from multi-modal data is applicable. In general, feature-level fusion (early fusion) and

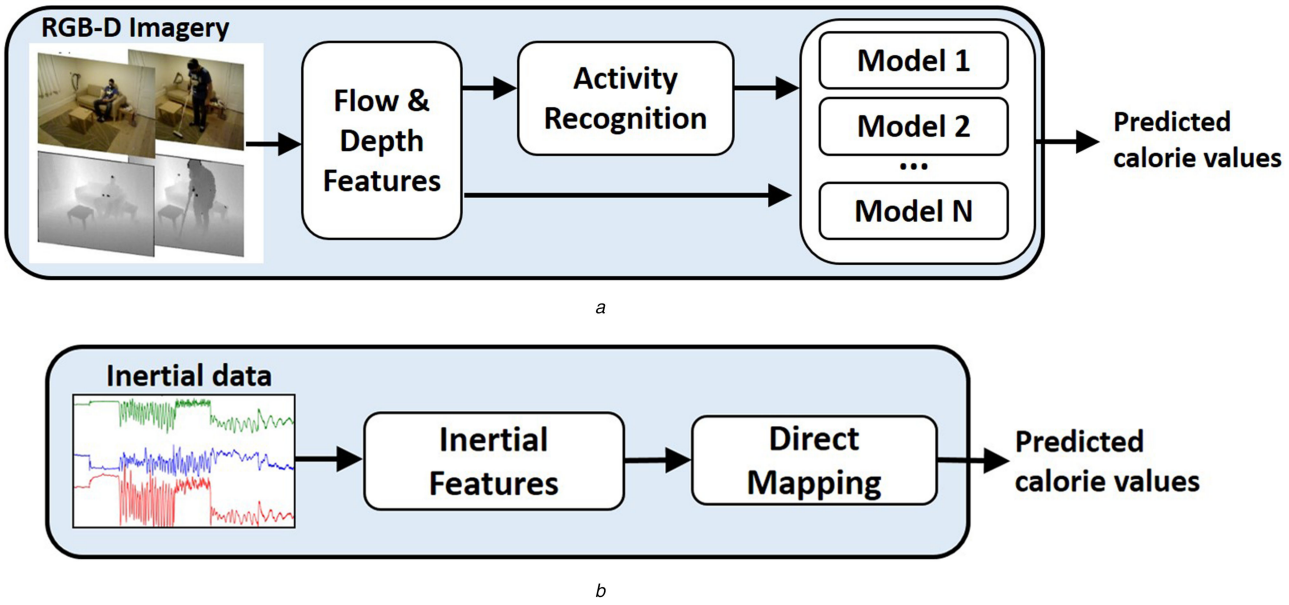


Fig. 2 *Overview of our visual-based and wearable-based frameworks*
(a) Visual-based framework. RGB-Depth videos are represented by a combination of flow and depth features. The proposed recurrent method then selects AS models which map to energy expenditure estimates, ***(b)*** Wearable-based framework. Inertial features are formed from two accelerometer sensor data, then features are mapped directly to calorie estimates via a monolithic classifier

decision-level fusion (late fusion) are the two approaches most often employed to fuse multiple modalities. Both early and late fusion strategies are explained in further detail in [46].

Feature-level fusion: This methodology involves carrying out fusion of features right after features are extracted from raw data. This scheme only requires one learning stage and allows to take advantage of mutual information from data. For instance, in [47], the depth and inertial sensor data were concatenated, then an Hidden Markov Model (HMM) classifier was employed for recognising basic gestures on the fused data. The results reveal significant improvements when the fusion scheme is applied compared to using each sensor individually. The work in [17] investigates the practical home-use of body-worn mobile phone inertial sensors together with an RGB-Depth camera to achieve monitoring of daily living scenarios. The results indicate that the vision-based approach significantly outperforms the wearable-based method, while fusion of both sensors slightly improves the performance further. Clearly, feature-level fusion can be applied effectively in practical settings; however, it may suffer from the 'curse of dimensionality'.

Decision-level fusion: This approach fuses the decisions made by individual classifiers, each of which corresponding to one sensor. Since decision information is of low complexity, the curse of dimensionality can effectively be targeted. In [48], for instance, a Bayesian co-boosting training framework combines multiple hidden Markov model classifiers of two modalities – a Kinect sensor and an inertial measurement unit. The result is the construction of a strong classifier for gesture recognition, which achieved the best performance in the multi-modal gesture recognition challenge. A real-time action recognition system in [49] uses Dempster–Shafer theory to combine the classification outcomes from a depth camera and several inertial sensors. A Bayesian model for sensor fusion is introduced in [16], which aims at addressing the challenges of fusion of heterogeneous sensor modalities in ambient assisted living.

Comparisons: In this work, we consider both fusion approaches and provide a direct comparison. In the feature-level fusion approach, features generated from the two modality sensors are merged before classification, and the decision-level fusion is performed by forming a linear combination of different classifiers using stacking regression [50] to improve overall accuracy. As outlined in the following section, our work attempts to use skeleton-independent, RGB-Depth-based vision, together with two wearable accelerometer devices to estimate calorific expenditure against a standardised calorimetry sensor COSMED-K4b2 based on gas exchange.

3 Method

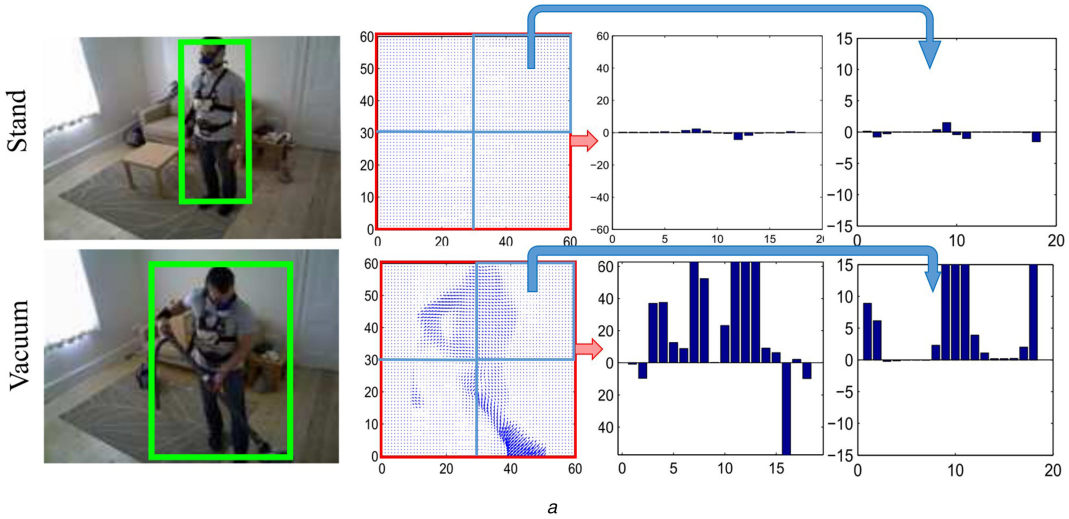
To describe our framework for estimating calorific expenditure, we initially introduce the methods for visual and wearable sensors separately, and then describe two approaches for their fusion.

Fig. 2*a* shows a flowchart of the visual method – mapping visual flow and depth features to calorie estimates using AS models. The method implements a cascaded and recurrent approach, which explicitly detects activities as an intermediate to select type-specific mapping functions for the final calorific estimation. Importantly, our setup as a video-based system is designed to reason about activities *first*, before estimating calorie expenditure via a set of models which are each separately trained for particular activities. In contrast to this, our direct mapping (DM) method designed for wearable sensor data directly maps inertial features to calorie estimates via a monolithic classifier. A flowchart of the wearable approach is shown in Fig. 2*b*. In our fusion system, we consider both feature-level and decision-level fusion of these two approaches. Finally, we compare these methods against a ground truth of gas-exchange measurements and off-the-shelf alternatives, that is manual mapping from activity classes to calorie estimates via METs lookup tables [2] as it is often applied in clinical practice today.

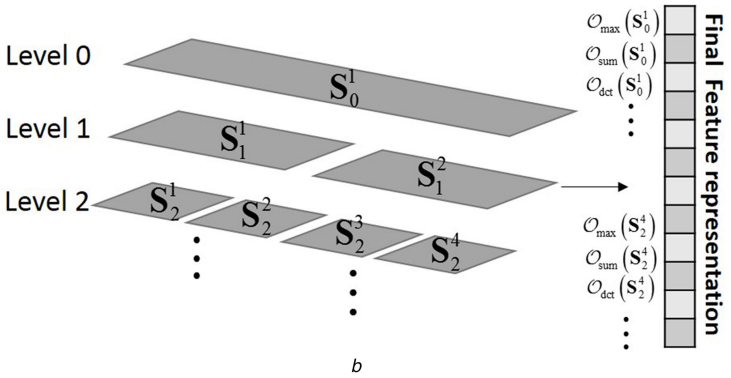
3.1 Visual features

We obtain RGB and depth imagery using an Asus Xtion. For each frame t, appearance and motion features are extracted, with the latter being computed with respect to the previous frame (level 0). A set of temporal filters is then applied to form higher level motion features (level 1). We extract features from within the bounding box returned by the OpenNI SDK [51] person detector and tracker, which allows to follow up to six persons in the camera view at the same time. To normalise the utilised image region due to varying heights of the subjects and their distance to the camera, the bounding box is scaled by fixing its longer side to $M = 60$ pixels, a size recognised as optimal for human action recognition [52], while maintaining aspect ratio. The scaled bounding box is then centred in a $M \times M$ square box and horizontally padded.

Motion feature encoding: Inspired by Tran and Sorokin [52], optical flow measurements are taken over the bounding box area and split into horizontal and vertical components. These are re-



a



b

Fig. 3 *Flow feature encoding via spatial pyramids and temporal pyramid pooling and its feature representation*

(a) Flow feature encoding via spatial pyramids. First row: limited motion while standing still. Second row: significant motion features when moving during vacuuming. First column: colour images with detected person. Second column: optical flow patterns. Third column: motion features at level 0. Last column: motion features from the top-right quadrants of the image at level 1 (at which the image is subdivided into four quadrants), ***(b)*** Temporal pyramid pooling and its feature representation. This schematic shows the temporal subdivision of data into various pyramidal levels (left) and the concatenation of resulting feature (e.g. max, sum, and dct) into a descriptor vector (right)

sampled to fit the normalised box and a median filter with kernel size 5×5 is applied to smooth the data. A spatial pyramid structure is used to form hierarchical features from this. Such partitioning of the image into an iteratively growing number of sub-regions increases discriminative power. The normalised bounding box is divided into a $n_g \times n_g$ non-overlapping grid, where n_g depends on the pyramid level, and the orientations of each grid cell are quantised into n_b bins. The parameters for our experiments are empirically determined as $n_b = 9$ and $n_g = 1$ and 2 for levels 0 and 1, respectively. Fig. 3*a* exemplifies optical flow patterns and their encoding in two different example activities.

Appearance feature encoding: We extract depth features by applying the histogram of oriented gradients feature on raw depth images [53] within the normalised bounding box. We then apply principal component analysis and keep the first 150 dimensions of this high-dimensional descriptor, which retains 95% of the total variance.

Pyramidal temporal pooling: Given the motion and appearance features extracted from each frame in a sequence of images, it is important to capture both short- and long-term temporal changes, and summarise them to represent the motion in the video. Pooled motion features were first presented in [43], even though designed for egocentric video analysis. We modify their pooling operator to make it more suitable for our data as follows – an illustration of the temporal pyramid structure and the process for pooling operations are shown in Fig. 3*b*. The time series data $\boldsymbol{S}$ can be represented as a set of time segments at level i as $\boldsymbol{S} = [\boldsymbol{S}_i^1, \ldots, \boldsymbol{S}_i^{2^i}]$. The final feature representation is a concatenation of multiple pooling operators applied to each time segment at each level. The time series data can also be explained as T per-frame feature vectors, such that $\boldsymbol{S} = \{S_1, \ldots, S_N\}$, $\boldsymbol{S} \in \mathbb{R}^{N \times T}$ for a video in matrix form, where N is the length of the per-frame feature vector, and T is the number of frames. A time series $S_n = [s_n(1), \ldots, s_n(T)]$ is the nth feature across $1, \ldots, T$ frames, where $s_n(t)$ denotes nth feature at frame t. A set of temporal filters with multiple pooling operators is applied to each time segment $[t_{\min}, t_{\max}]$ and produces a single feature vector for each segment via concatenation. We use two conventional pooling operators, max pooling and sum pooling, as well as frequency domain pooling. They are defined as

$$\mathcal{O}_{\max}(S_n) = \max_{t = t_{\min} \cdots t_{\max}} s_n(t) \quad \text{and} \quad \mathcal{O}_{\text{sum}}(S_n) = \sum_{t = t_{\min}}^{t_{\max}} s_n(t) \tag{1}$$

Frequency domain pooling is used to represent the time series S_n in the frequency domain by the discrete cosine transform (dct), where the pooling operator takes the absolute value of the j lowest frequency components of the frequency coefficients D

$$\mathcal{O}_{\text{dct}}(S_n) = \left| M_{1:j} S_n \right| \tag{2}$$

where $\boldsymbol{M}$ is the discrete cosine transformation matrix.

3.2 Inertial features

Raw time series data from accelerometers is measured as $[\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}]$ vectors, where each column corresponds to acceleration in orthogonal spatial dimensions. Fig. 1 illustrates the raw accelerometer data collected from one wearable device for various actions. From the raw data the pooled motion features are formed from each of the three axes for each device. Abstracting short-term and long-term changes in the inertial feature descriptor is essential; it is particularly useful for modelling the level of activity intensity

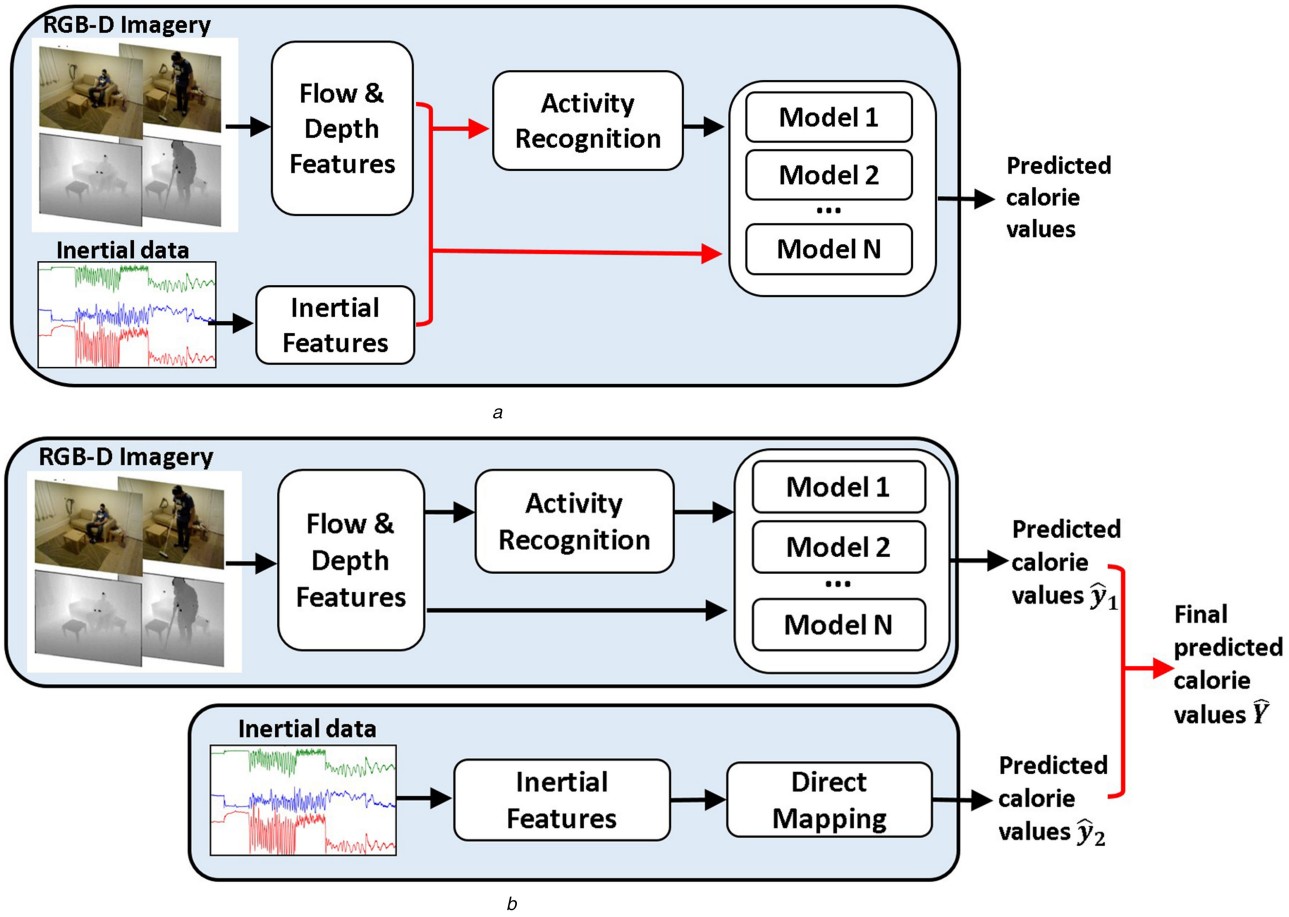


Fig. 4 *Fusion approaches overview*

(a) Feature-level fusion framework. The features from visual and inertial sensors are concatenated to form a monolithic input into activity recognition and AS models, ***(b)*** Decision-level fusion framework. Calorie values are predicted individually by the different sensor modalities, and then combined using a regression method to form final calorie estimates

changes. Thus, we apply three pooling operators (max pooling, sum pooling, and frequency domain pooling) to the inertial data.

3.3 Learning and recurrency

Energy expenditure estimation can be formulated as a sequential and supervised regression problem. We train a support vector regressor to predict calorie values from given features over a training set. The sliding window method is used to map each input window of width w to an individual output value y_t. The window contains the current and the previous $w-1$ observations. The window feature is represented by temporal pooling from the time series $\boldsymbol{S} = \{S^{t-w+1}, \ldots, S^t\}$.

We note that energy values for a particular time are highly dependent on the energy expenditure history. In our system, these are most directly expressed by previous calorific predictions during operation. Thus, employing recurrent sliding windows offers an option to not only use the features within a window, but also take the most recent d predictions $\{\hat{y}^{t-d}, \ldots, \hat{y}^{t-1}\}$ into consideration to help predict y^t. During learning, as suggested in [54], the ground truth labels in the training set are used in place of recurrent values.

3.4 Fusion approach

Both feature-level and decision-level fusion are considered in our work.

Feature-level fusion: This is an early fusion approach, for which all features from all modalities are concatenated together, and employed as a single unified feature stream to the learning components. Given visual features in d_1-dimensional feature space $\boldsymbol{S}_v \in \mathbb{R}^{d_1}$ and accelerometer features in d_2-dimensional feature space $\boldsymbol{S}_a \in \mathbb{R}^{d_2}$, the fused feature set can be represented as $\boldsymbol{S} \in \mathbb{R}^{d_1+d_2}$, where the feature set is constructed as $\boldsymbol{S} = (\boldsymbol{S}_v, \boldsymbol{S}_a)$. The fused feature vector is then used as input to the classifiers of the system. Fig. 4*a* shows a flowchart of this feature-level fusion approach.

Decision-level fusion: In this approach, a collection of models are learned, and the predictions are combined together only at the last stage to form the final decision. We apply the decision-level fusion via a stacking regression method, which forms linear combinations of different classifiers to improve overall estimation accuracy.

Consider that there are K predicted values $\hat{y}_1, \ldots, \hat{y}_K$ estimated from each regressor individually. Then, the final predictor value $\hat{Y}(\boldsymbol{S})$ can be represented as a linear combination of a set of predicted values with different weighting coefficients, constructed as

$$\hat{Y}(\boldsymbol{S}) = \sum_{k=1}^{K} \alpha_k \hat{y}_k(\boldsymbol{S}) \tag{3}$$

Given a set of training data $\{(\boldsymbol{S}^1, y^1), \ldots, (\boldsymbol{S}^T, y^T)\}$ with T training samples, where each $\boldsymbol{S}^t$ is an input vector, the goal is to minimise the distance of the ground truth y^t and the predicted values $\hat{Y}^t(\boldsymbol{S})$ from the combined regressor. This optimised distance can be obtained by

$$\arg\min_{\alpha_k} \sum_{t=1}^{T} \left(y^t - \sum_{k=1}^{K} \alpha_k \hat{y}_k^t(\boldsymbol{S}^t) \right)^2, \tag{4}$$

with the constraints $0 \le \alpha_k \le 1$, $k = 1, \ldots, K$. The resulting combined predicted value $\sum_{k=1}^{K} \alpha_k \hat{y}_k(\boldsymbol{S})$ is then used as prediction. Fig. 4*b* shows a flowchart of this decision-level fusion approach.

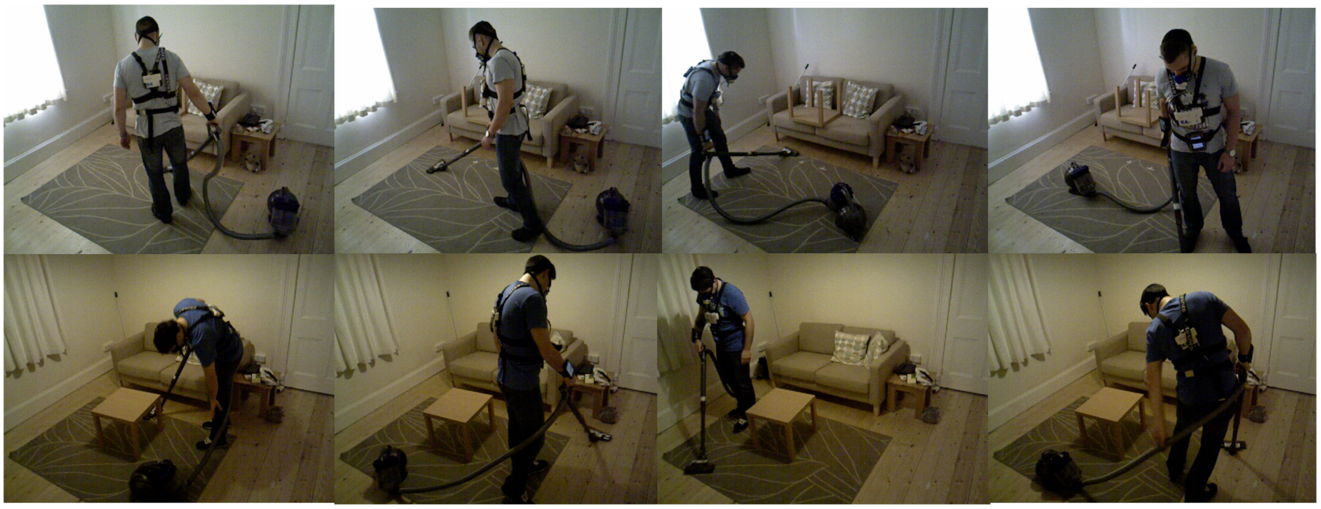


Fig. 5 *Example poses from the activity 'vacuuming'. It can be seen that the sequences contain a large variety of body positions, view-points, and distances naturally associated with the action. Two example sequences are captured in daytime and nighttime, respectively, indicating different lighting conditions*

Table 1 Activities, their associated MET values, and the levels of activity intensity

Intensity	Activity	MET value
light	*sit still*	1.3
	stand still	1.3
	lying down	1.3
	reading	1.5
light+	*walking*	2.0
	wiping table	2.3
	cleaning floor stain	3.0
moderate	*vacuuming*	3.3
	sweeping floor	3.3
	upper body exercise	4.0
	stretch	5.0

4 Experimental results

4.1 Dataset and ground truth

We introduce the *SPHERE_RGBD + Inertial_calorie* dataset for human calorific expenditure estimation, comprising RGB-Depth and inertial sensor data captured in a real living environment. The ground truth was captured by the COSMED K4b2 portable metabolic measurement system. The dataset was generated over 20 sessions by 10 subjects with varying anthropometric measurements. Participants were seven males and three females, with mean age of 27.2 ± 3.8 years, mean weight of 72.3 ± 15.0 kg, mean height of 173.6 ± 9.8, mean body mass index of 23.7 ± 2.8. Ethics approval was obtained and each participant signed a consent form agreeing to share their data for research purposes. The dataset contains up to 11 activity categories per session, and totalling around 10 h recording time. The activities were captured in daily-living scenarios containing a variety of body positions, view-points, and distances naturally associated with the various actions performed. Fig. 5 shows frames from the *vaccuming* activity depicting this variety. It is also shown that the sequences are captured in different time of the day which contains various lighting conditions. All the activities, the associated intensity categories, and MET values are shown in Table 1. In addition, Table 2 lists the number of frames for each action and sequence [Some actions in certain sequences are missing due to various reasons (hence they have 0 frames), for example *exercise* is missing in sequences 7 and 17 as the participants had difficulty in performing the action.].

Colour and depth images were acquired at a rate of 30 Hz. The accelerometer data was captured at about 100 Hz and sampled down to 30 Hz, a frequency recognised as optimal for human action recognition [55]. The calorimeter gives readings per breath, which occurs approximately every 3 s. To model transitions better between activity levels, we consider the nine different combinations of the three activity intensities (light, light+, moderate) in the design of each session.

Fig. 1 shows a detailed example of calorimeter readings and associated sample RGB images from the dataset (top) and the accelerometer data reading (bottom). The raw breath data is noisy (in red). We apply an average filter with a span of ∼20 breaths (in blue). The participants were asked to perform the activities based on their own living habits without any extra instructions.

4.2 Parameter settings

In our experiments, we use non-linear Support Vector Machines (SVMs) with radial basis function kernels for activity classification and a linear support vector regressor for energy expenditure prediction. The libsvm [56] implementation was used. We perform a grid search algorithm to estimate the hyper-parameters of the SVM. To test our individual-independent approach, we implement leave-one-subject-out cross-validation on the dataset in which each subject's data are tested in turn using models trained with all other subject data combined. This process iterates through all subjects, and the average testing error and standard deviation of all iterations are reported. We use the normalised root-mean-squared error (normalised RMSE) as a standard evaluation metric to facilitate the comparison between data with different scales for the deviation of estimated calorie values from the ground truth.

4.3 Evaluation of individual modalities

We start with tests on each sensor type (visual and inertial), and compare their performance in situations when used independently.

Temporal window size: The accuracy of predicted calorie values is linked to the window of previous information utilised for making the prediction. In a first experiment, we look at the relation

Table 2 Number of frames per sequence and action in *SPHERE_RGBD + Inertial_calorie* dataset

seq ID	*stand*	*sit*	*walk*	*wipe*	*vacuum*	*sweep*	*lying*	*exercise*	*stretch*	*clean*	*read*	overall
1	3739	5230	5509	5117	5447	5113	5142	2789	2421	5072	6092	51,671
2	3546	5388	5304	5244	5177	5091	5020	2545	2566	5112	5311	50,304
3	3876	5682	6093	4745	5151	4946	5360	2297	2318	5317	5095	50,880
4	3948	5211	5470	5182	5025	4813	3784	0	0	0	0	33,433
5	3239	5133	5670	4878	5436	4327	0	0	0	0	0	28,683
6	3812	5294	5889	5023	5015	4841	4878	3205	1778	4738	5070	49,543
7	3796	5239	11,602	6684	4432	4561	3590	0	5474	5837	5337	56,552
8	3951	5257	5412	5244	4886	1005	0	0	0	0	0	25,755
9	4128	5568	5091	5195	4651	3619	4891	2458	2483	5700	5298	49,082
10	3649	5202	5317	5354	4651	4990	5030	2330	1669	3933	5337	47,462
11	4367	4901	5503	5133	5006	4761	0	0	0	0	0	29,671
12	3697	5270	5618	5010	5107	4988	4991	2891	2335	5299	5412	50,618
13	4250	5936	4644	5162	5259	4517	4944	899	2839	5495	6206	50,151
14	4263	4732	5370	5150	4769	4847	4861	3008	3224	4366	5765	50,355
15	3457	5784	4789	4745	5159	4911	0	0	0	0	0	28,845
16	3919	5466	5062	5308	2716	0	0	0	0	0	0	22,471
17	3613	5343	5432	4914	5032	4461	4979	0	5063	5902	4873	49,612
18	3715	5340	5422	5013	5893	4743	4517	1977	2793	5948	5012	50,373
19	4521	5434	5787	4740	5015	4459	5480	3174	2707	4803	6342	52,462
20	4040	5255	5597	5472	5309	4551	4926	1797	1691	6356	6443	51,437

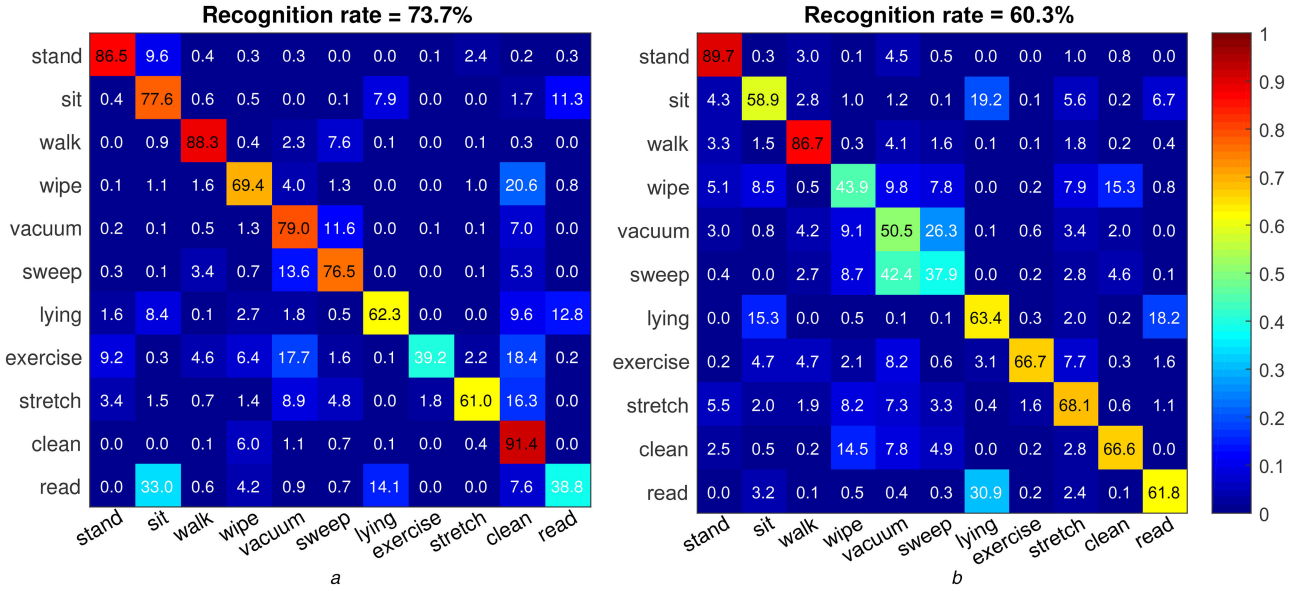


Fig. 6 *Recognition confusion matrices from the best activity recognition results corresponding to the use of* ***(a)*** Visual sensor, ***(b)*** Inertial sensors

Table 3 Vision-based prediction results. Activity recognition rate (%) and calorific expenditure prediction error (normalised RMSE) with different window lengths *w*, stated in seconds. The best results for each activity are in bold

w		*stand*	*sit*	*walk*	*wipe*	*vacuum*	*sweep*	*lying*	*exercise*	*stretch*	*clean*	*read*	overall
7.5	activity	83.5	75.2	**90.5**	70.2	79.7	73.4	58.6	36.6	54.1	**93.8**	**39.9**	72.3
	calorie	0.84	0.70	**0.29**	0.52	**0.31**	0.41	0.74	0.63	0.52	0.41	0.67	0.58
15	activity	**86.5**	77.6	88.3	69.4	79.0	**76.5**	**62.3**	39.2	**61.1**	91.4	38.9	**73.7**
	calorie	0.83	0.66	0.30	0.46	0.34	0.45	0.74	0.66	0.54	0.40	0.64	0.53
30	activity	85.0	79.1	89.4	**71.9**	**81.1**	75.2	54.3	**40.3**	57.8	90.4	36.8	71.1
	calorie	0.73	0.52	**0.30**	**0.41**	0.36	0.41	0.66	0.55	0.54	**0.37**	0.54	0.49
60	activity	81.1	**79.7**	85.1	66.0	77.2	72.9	33.0	29.3	52.7	90.0	35.9	68.2
	calorie	**0.54**	**0.45**	0.32	0.44	**0.34**	**0.39**	**0.58**	**0.42**	**0.52**	0.38	**0.50**	**0.45**

between window length on the one hand, and activity recognition and calorie prediction errors on the other. All sequences are tested with temporal windows of $w = \{7.5, 15, 30, 60\}$ s. Tables 3 and 4 illustrate the activity recognition rates and the average normalised RMSEs for calorie prediction at different window length w using visual and accelerometer data, respectively. It can be seen that, in both modalities, the best performance across the set for recognising activities is achieved when a relatively small size of window is applied. The confusion matrices corresponding to the use of visual and inertial sensors are depicted in Fig. 6.

In a second experiment, we test how the estimated calorie value is influenced by the performance of action recognition. The results

Table 4 Inertial-based prediction results. Activity recognition rate (%) and calorific expenditure prediction error (normalised RMSE) with different window lengths *w*, stated in seconds. The best results for each activity are in bold

w		*stand*	*sit*	*walk*	*wipe*	*vacuum*	*sweep*	*lying*	*exercise*	*stretch*	*clean*	*read*	overall
7.5	activity	**91.1**	58.7	78.1	43.6	50.1	**48.2**	66.0	51.8	65.2	59.8	48.0	60.1
	calorie	0.54	0.50	0.49	0.52	0.51	0.41	0.64	0.83	0.72	0.51	0.77	0.63
15	activity	89.7	**58.9**	**86.7**	43.9	**50.5**	37.9	63.4	**66.7**	**68.1**	**66.6**	**61.8**	**60.3**
	calorie	0.50	0.51	**0.41**	**0.45**	**0.46**	0.38	**0.53**	0.80	0.68	0.44	0.73	0.59
30	activity	88.7	56.9	83.9	46.6	48.7	36.4	69.4	51.8	64.1	61.0	52.1	58.0
	calorie	0.46	0.49	0.48	0.46	0.52	**0.35**	0.54	0.77	0.60	0.43	0.71	0.56
60	activity	90.8	55.3	74.8	**52.3**	49.6	45.7	**73.1**	44.9	59.0	64.2	56.0	58.3
	calorie	**0.41**	**0.41**	0.51	0.50	0.47	0.40	0.57	**0.64**	**0.55**	**0.41**	**0.71**	**0.54**

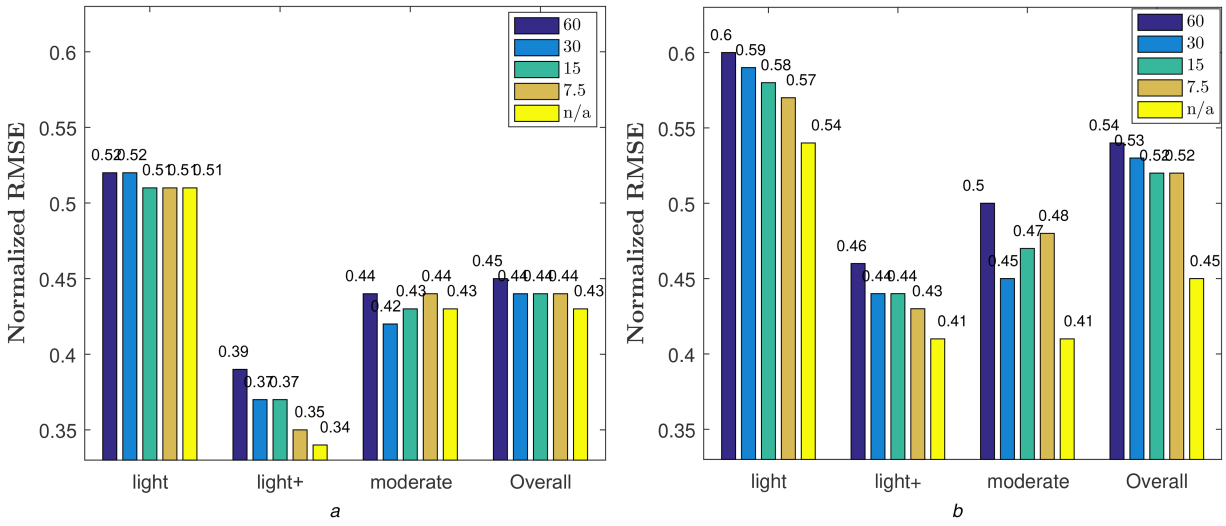


Fig. 7 *Prediction accuracy of calorific expenditure. Average calorie prediction errors (normalised RMSE) when ground truth labels are used to select the AS model (in yellow), and when action recognition is employed at different window length using*
(a) Visual sensor, ***(b)*** Two accelerometer sensors

Table 5 Calorific expenditure prediction error (normalised RMSE) using the visual sensor when ground truth labels are used to select the AS model (top row) and when action recognition is employed at different window lengths. The best results for each activity are in bold

calorie *w*	activity *w*	*stand*	*sit*	*walk*	*wipe*	*vacuum*	*sweep*	*lying*	*exercise*	*stretch*	*clean*	*read*	overall
60	n/a	**0.40**	0.45	**0.28**	**0.35**	0.32	0.38	**0.55**	**0.36**	0.55	0.36	0.50	**0.43**
	7.5	0.51	**0.43**	**0.28**	0.38	0.33	0.38	0.58	0.43	0.55	**0.35**	**0.43**	0.45
	15	0.41	**0.43**	0.30	0.41	0.32	0.39	0.57	0.45	0.54	0.36	0.44	0.44
	30	0.47	0.44	0.30	0.42	**0.31**	**0.37**	0.56	0.44	0.53	0.37	0.46	0.44
	60	0.54	0.45	0.32	0.44	0.34	0.39	0.58	0.42	**0.52**	0.38	0.50	0.45

Table 6 Calorific expenditure prediction error (normalised RMSE) using the inertial sensors when ground truth labels are used to select the AS model (top row), and when action recognition is employed at different window lengths. The best results for each activity are in bold

calorie *w*	activity *w*	*stand*	*sit*	*walk*	*wipe*	*vacuum*	*sweep*	*lying*	*exercise*	*stretch*	*clean*	*read*	overall
60	n/a	0.47	**0.41**	**0.33**	**0.26**	**0.37**	0.45	**0.46**	**0.57**	0.47	0.40	**0.50**	**0.45**
	7.5	0.51	0.47	0.41	0.45	0.39	**0.40**	0.62	0.72	0.52	0.42	0.63	0.52
	15	0.53	0.57	0.40	0.48	0.44	0.43	0.57	0.72	**0.46**	**0.39**	0.57	0.52
	30	0.48	0.45	0.46	0.50	0.43	0.43	0.57	0.68	0.55	0.41	0.58	0.53
	60	**0.41**	**0.41**	0.51	0.50	0.47	**0.40**	0.57	0.64	0.55	0.41	0.71	0.54

for each activity using visual and inertial sensors are listed in Tables 5 and 6, respectively. We first test a system in which the ground truth labels are used to select the AS model for calorie prediction (top rows in Tables 5 and 6). We then compare actual action recognition at varying window lengths $w = \{7.5, 15, 30, 60\}$ s. In all cases, we use a fixed window length $w = 60$ s for calorific expenditure estimation to focus on the effect of varying action recognition quality. As expected, it can be observed that for most activities, the calorie estimation error is smallest when there is no activity recognition error (top rows in Tables 5 and 6). For a more detailed visualisation, we also show the results in Fig. 7; the 11 actions are grouped into three clusters based on their intensity level (see Table 1). The figure summarises the calorie prediction error for different intensities and action recognition rates using the visual system and the inertial system, respectively.

Model comparison: As just observed, activity recognition accuracy affects the calorie prediction results. To determine if the AS model provides a predictive advantage, we compare the estimation performance of the AS approach against the DM approach for each sensor modality. For both sensor systems, we select a fixed window length of $w = 60$ s to analyse performance for calorie value prediction in both DM and AS, and $w = 15$ s for

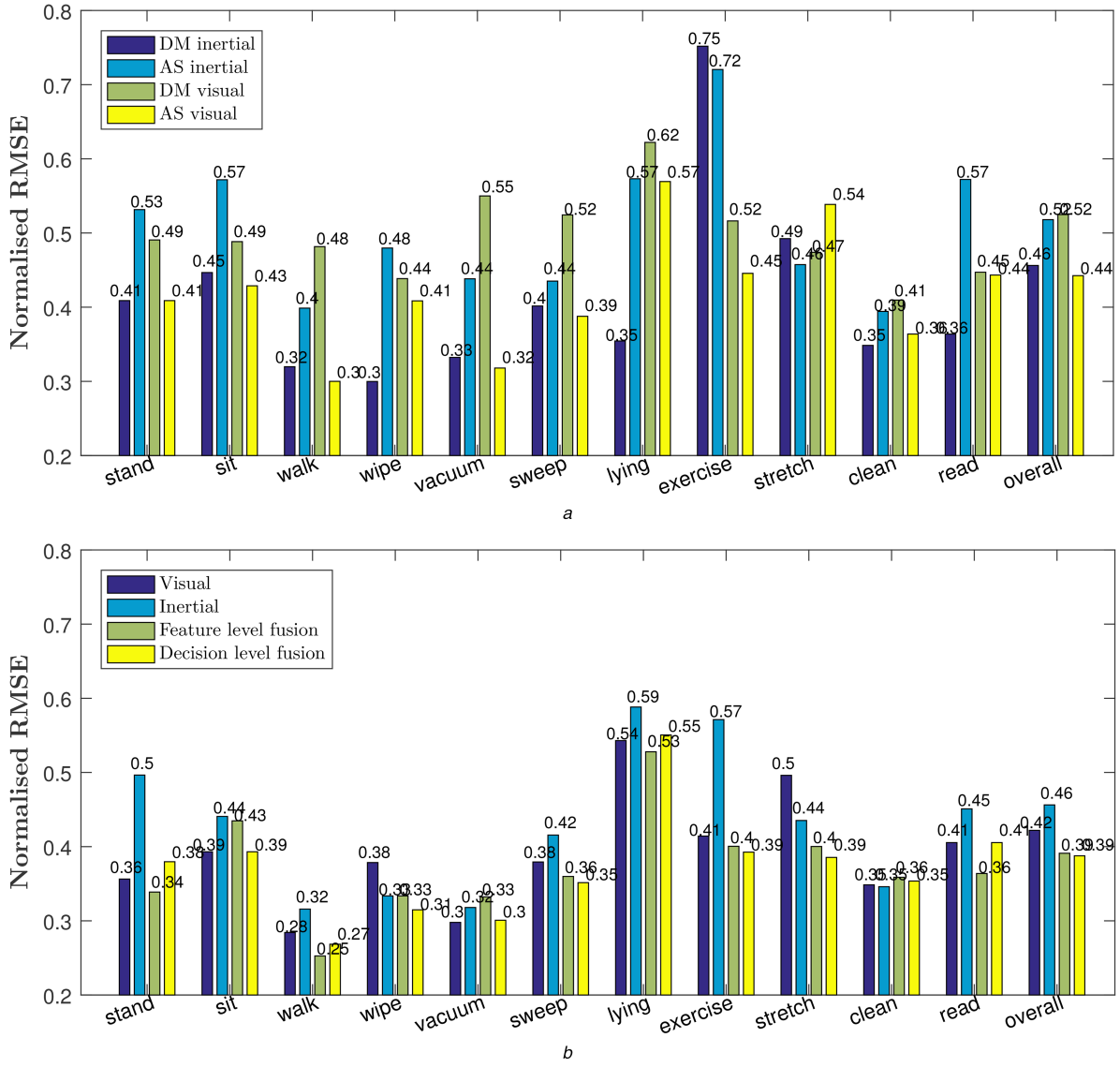


Fig. 8 *Average calorie prediction errors*

(a) Average calorie prediction errors (normalised RMSE) of DM and AS approaches using visual and inertial sensors, respectively, ***(b)*** Average calorie prediction errors (normalised RMSE) of using visual sensor only (visual), inertial sensors only (inertial), and two sensor fusion approaches

Table 7 Average calorific expenditure prediction errors (normalised RMSE) for each activity with different learning approaches. The best results for each activity are in bold

	w	*stand*	*sit*	*walk*	*wipe*	*vacuum*	*sweep*	*lying*	*exercise*	*stretch*	*clean*	*read*	overall
visual	baseline	0.41	0.43	0.30	0.41	0.32	0.39	0.57	0.45	0.54	0.36	0.44	0.44
	recurrent1	0.36	**0.39**	**0.28**	**0.38**	**0.30**	**0.38**	**0.54**	0.41	0.50	**0.35**	**0.41**	**0.42**
	recurrent2	**0.35**	0.49	0.31	0.52	0.59	0.53	0.56	**0.38**	**0.46**	0.42	0.45	0.52
inertial	baseline	**0.41**	**0.45**	**0.32**	**0.30**	**0.33**	**0.40**	**0.35**	**0.75**	**0.49**	**0.35**	**0.36**	**0.46**
	recurrent1	**0.41**	0.54	0.37	0.32	0.34	0.41	0.45	0.78	**0.49**	0.36	0.44	0.49
	recurrent2	0.50	1.06	0.63	0.49	0.45	0.49	0.92	0.95	0.52	0.44	0.85	0.68

activity recognition in AS. The results are shown in Fig. 8*a* for each activity. It can be seen that for the visual-based system, the AS provides best prediction results overall and significantly outperforms DM in most activities. For the inertial-based system, the error associated with AS is significantly higher compared with DM. This is in part due to poor activity recognition results in an inertial measurement setup, which effectively leads to using wrong models to estimate calorie values.

Evaluation of a recurrent system layout: To evaluate the use of recurrency, we set the AS method using the sliding window technique as our baseline method for the vision-based comparison, and the DM method for inertial-based comparison. We now introduce two methods, which are based on recurrent sliding window approaches. The first method (Recurrent1) uses the most recent predictions of the baseline method as input together with visual/inertial features to predict current calorie value. Thus, it implements indirect recurrency utilising the predicted values from the baseline as recent predictions. The second method (Recurrent2) implements full recurrency, i.e. it uses its own output as recurrent input together with visual/inertial features.

Table 7 shows the effect of using recurrent information, with the best results for each activity highlighted. In general, the full recurrency model, Recurrent2, suffers from drift and produces the worst results for half of the activities and also overall. When the visual sensor is used, indirect recurrency, Recurrent1, outperforms the other approaches at an average normalised RMSE of 0.42, while in inertial-based systems indirect recurrency increases the estimation error by 7% comparing to its baseline.

Table 8 Ground truth and predicted calorie values in total per sequence and its accuracy and correlation

	Prediction (calories)					Accuracy %				Correlation			
Sequence	GT	Visual	Inertial	Fusion	MET	Visual	Inertial	Fusion	MET	Visual	Inertial	Fusion	MET
1	59	71	76	63	76	80.2	71.1	**93.9**	71.3	0.83	0.40	**0.88**	0.66
2	89	80	83	95	78	90.3	93.3	**93.5**	88.2	**0.85**	0.67	0.76	0.57
3	74	81	83	78	69	90.1	88.6	**94.1**	92.7	**0.84**	0.73	0.83	0.63
4	79	48	52	56	43	60.4	66.1	**71.5**	55.0	0.87	**0.89**	0.87	0.78
5	37	39	50	40	28	**98.6**	63.1	91.5	77.6	**0.90**	0.82	0.88	0.77
6	89	86	88	87	107	94.3	**99.0**	90.6	98.1	0.82	0.73	**0.83**	0.63
7	101	96	94	109	114	**95.3**	92.7	92.0	87.6	0.61	0.65	**0.66**	0.61
8	39	42	40	44	35	91.9	**97.2**	88.6	84.4	**0.93**	0.88	0.91	0.57
9	82	76	81	90	94	92.8	**99.3**	89.6	85.3	**0.86**	0.62	**0.86**	0.71
10	49	68	70	64	76	61.5	57.5	**68.6**	45.2	0.54	0.31	**0.55**	0.42
11	28	38	45	41	38	65.5	42.2	56.6	**66.8**	0.64	0.48	**0.66**	0.56
12	98	88	90	104	79	90.3	91.4	**93.8**	80.8	0.56	0.48	0.62	**0.66**
13	56	66	82	60	77	82.3	53.6	**91.9**	62.7	**0.78**	0.66	0.82	0.62
14	141	84	86	104	74	57.4	60.8	**73.8**	52.8	0.86	0.75	**0.87**	0.60
15	40	41	49	42	30	**98.9**	76.3	93.0	74.5	0.94	0.88	**0.96**	0.94
16	29	31	34	32	38	**97.3**	79.9	88.0	69.1	**0.88**	0.85	0.84	0.81
17	81	85	80	88	100	94.2	**99.5**	90.6	76.0	**0.74**	0.69	0.72	0.70
18	65	86	78	69	94	69.3	79.4	**93.1**	54.7	0.83	0.64	**0.85**	0.48
19	92	89	84	99	101	**94.2**	91.0	93.0	90.6	0.75	0.56	**0.77**	0.72
20	63	83	79	77	86	66.9	74.0	**77.0**	64.4	0.81	0.66	**0.86**	0.41
average	—	—	—	—	—	82.9	78.8	**86.6**	73.7	0.80	0.67	**0.81**	0.64

The best results for each sequence are in bold.

4.4 Comparing sensor fusion approaches

Having tested the two modalities individually, we now study modality fusion approaches against the use of individual sensor systems, and also compare against the MET lookup table method for completeness.

In feature level fusion, we apply our AS approach with $w = 15$ s for activity recognition and $w = 60$ s for calorific expenditure estimation. In decision level fusion, we again use the most suitable model for each sensor data to fuse, which is the AS approach for visual sensor data, and DM approach for inertial sensor data. The estimation performance of the two fusion approaches is compared with the performance of each sensor modality individually, as shown in Fig. 8*b*. It can be seen that both fusion approaches on average outperform unimodal prediction. In particular, by combining the features from the visual and the inertial data, the overall prediction error decreases from 0.46 (inertial sensors alone) and 0.42 (visual sensor alone) to 0.39. The calorie prediction accuracy for most activities is improved when using fusion approaches. We also observed that the two fusion frameworks achieve similar performance.

Finally, we present the results produced by MET, which is commonly used by clinicians and physiotherapists, and compare our proposed methods against it. It assumes N clusters of activity $\mathbf{A} = \{A_1, A_2, \ldots, A_N\}$ are known. A MET value is assigned to each cluster, together with anthropometric characteristics of individuals. The amount of AS energy expended can then be estimated as energy $= 0.0175(\text{kcal/kg/min}) \times \text{weight (kg)} \times \text{MET}$ values [2]. Here, we use the ground truth labels to select activities to keep this procedure identical to the commonly used manual estimate. Table 8 presents the detailed results for each sequence. The accuracy is calculated over the total calorie expended in each recording session. We also measure the correlation between the ground truth and the observed values [Note that the total calorie values for sequence 4, 5, 8, 11, 15, and 16 are relatively low due to shorter sequences.]. We can see that the fusion of visual and inertial sensors achieves higher accuracy and correlation in more sequences than the MET model or unimodal approaches, and obtains better rates on average, which points towards an advantage of using visual-inertial setups for the task of calorific expenditure prediction.

5 Conclusion and future directions

We have presented a system for calorific expenditure estimation using data from two different modality sensors, a RGB-Depth camera sensor, and wearable inertial sensors (accelerometers). We have demonstrated the effectiveness of the fusion approach through a comprehensive comparative study with single modality setups and widely used METs prediction. The proposed fusion system used pooled spatial and temporal pyramids of visual and accelerometer features, which subsequently are fed in both early and late fusion approaches. To test the methodology, we introduced the challenging *SPHERE_RGBD + Inertial_calorie* dataset, which covers a wide variety of home-based human activities. The proposed fusion method demonstrates its ability to outperform the METs estimation approach and the use of single modality sensors. The focus of the paper has been on presenting a system for estimating calorific expenditure from combined visual and accelerometer sensors, where the purpose of the study has been to show that the fusion of both modalities improves the estimates beyond the accuracy of single modality, and the proposed system outperforms manual metabolic lookup table based methods – the main measure used in clinical practice today. We acknowledge that applying more advanced fusion approaches and different feature representations may improve the performance further. Possible future directions include introducing deep learning models and investigating advanced data fusion methodologies for different modality sensors. We hope this work, and the new dataset, will establish a baseline for future research in the area.

6 Acknowledgment

This work was performed under the SPHERE IRC project funded by the UK Engineering and Physical Sciences Research Council, Grant EP/K031910/1. The data from this study is available by request from the University of Bristol research data repository via http://www.irc-sphere.ac.uk/work-package-2/calorie or http://doi.org/cc5k.

7 References

[1] Samitz, G., Egger, M., Zwahlen, M.: 'Domains of physical activity and all-cause mortality: systematic review and dose–response meta-analysis of cohort studies', *Int. J. Epidemiol.*, 2011, **40**, (5), pp. 1382–1400

[2] Ainsworth, B.E., William, L.H., Melicia, C.W., *et al.*: 'Compendium of physical activities: an update of activity codes and met intensities', *Med. Sci. Sports Exercise*, 2000, **32**, (9), pp. 498–504

[3] Ravussin, E., Lillioja, S., Anderson, T., *et al.*: 'Determinants of 24-hour energy expenditure in man. Methods and results using a respiratory chamber', *J. Clin. Invest.*, 1986, **78**, (6), p. 1568

[4] Cosmed K4b2. Available at http://www.cosmed.com/

[5] Chen, C., Jafari, R., Kehtarnavaz, N.: 'Improving human action recognition using fusion of depth camera and inertial sensors', *IEEE Trans. Hum.-Mach. Syst.*, 2015, **45**, (1), pp. 51–61

[6] Gjoreski, H., Kaluža, B., Gams, M., *et al.*: 'Context-based ensemble method for human energy expenditure estimation', *Appl. Soft Comput.*, 2015, **37**, pp. 960–970

[7] Aggarwal, J., Xia, L.: 'Human activity recognition from 3D data: a review', *Pattern Recognit. Lett.*, 2014, **48**, pp. 70–80

[8] Zhu, N., Diethe, T., Camplani, M., *et al.*: 'Bridging e-health and the internet of things: the sphere project', *IEEE Intell. Syst.*, 2015, **30**, (4), pp. 39–46

[9] Woznowski, P., Fafoutis, X., Song, T., *et al.*: 'A multi-modal sensor infrastructure for healthcare in a residential environment'. IEEE Int. Conf. on Communication Workshop (ICCW), 2015, 2015, pp. 271–277

[10] Planinc, R., Chaaraoui, A.A., Kampel, M., *et al.*: 'Computer vision for active and assisted living', *Act. Assist. Living, Technol. Appl.*, 2016, **57**, pp. 1–23

[11] Pitsikalis, V., Katsamanis, A., Theodorakis, S., *et al.*: 'Multimodal gesture recognition via multiple hypotheses rescoring', *J. Mach. Learn. Res.*, 2015, **16**, (1), pp. 255–284

[12] Leutenegger, S., Lynen, S., Bosse, M., *et al.*: 'Keyframe-based visual–inertial odometry using nonlinear optimization', *Int. J. Robot. Res.*, 2015, **34**, (3), pp. 314–334

[13] Fang, W., Zheng, L., Deng, H., *et al.*: 'Real-time motion tracking for mobile augmented/virtual reality using adaptive visual-inertial fusion', *Sensors*, 2017, **17**, (5), p. 1037

[14] Gasparrini, S., Cippitelli, E., Gambi, E., *et al.*: 'Proposal and experimental evaluation of fall detection solution based on wearable and depth data fusion', *ICT Innov.*, 2015, **1** (2016), pp. 99–108

[15] Stein, S., McKenna, S.J.: 'Combining embedded accelerometers with computer vision for recognizing food preparation activities'. Proc. of the 2013 ACM Int. Joint Conf. on Pervasive and Ubiquitous Computing, 2013, pp. 729–738

[16] Diethe, T., Twomey, N., Kull, M., *et al.*: 'Probabilistic sensor fusion for ambient assisted living', arXiv preprint arXiv:1702.01209

[17] Tao, L., Burghardt, T., Hannuna, S., *et al.*: 'A comparative home activity monitoring study using visual and inertial sensors'. IEEE Int. Conf. on E-Health Networking, Application and Services, 2015, pp. 644–647

[18] Tao, L., Burghardt, T., Mirmehdi, M., *et al.*: 'Real-time estimation of physical activity intensity for daily living'. 2nd IET Int. Conf. on Technologies for Active and Assisted Living, 2016, pp. 11–16

[19] Tao, L., Burghardt, T., Mirmehdi, M., *et al.*: 'Calorie counter: RGB-depth visual estimation of energy expenditure at home'. Asian Conf. on Computer Vision, Workshop on Assistive Vision, 2016, 0

[20] Edgcomb, A., Vahid, F.: 'Estimating daily energy expenditure from video for assistive monitoring'. Int. Conf. on Healthcare Informatics, 2013, pp. 184–191

[21] Tsou, P.-F., Wu, C.-C.: 'Estimation of calories consumption for aerobics using kinect based skeleton tracking'. Int. Conf. on Systems, Man, and Cybernetics, 2015, pp. 1221–1226

[22] Lara, O.D., Labrador, M.A.: 'A survey on human activity recognition using wearable sensors', *IEEE Commun. Surv. Tutor.*, 2013, **15**, (3), pp. 1192–1209

[23] Igual, R., Medrano, C., Plaza, I.: 'Challenges, issues and trends in fall detection systems', *Biomed. Eng. Online*, 2013, **12**, (1), p. 1

[24] Qudah, I., Leijdekkers, P., Gay, V.: 'Using mobile phones to improve medication compliance and awareness for cardiac patients'. Proc. of the 3rd Int. Conf. on PErvasive Technologies Related to Assistive Environments, 2010, vol. **36**

[25] Bennett, T.R., Wu, J., Kehtarnavaz, N., *et al.*: 'Inertial measurement unitbased wearable computers for assisted living applications: a signal processing perspective', *IEEE Signal Process. Mag.*, 2016, **33**, (2), pp. 28–35

[26] Ravi, N., Dandekar, N., Mysore, P., *et al.*: 'Activity recognition from accelerometer data'. AAAI, 2005, vol. **5**, pp. 1541–1546

[27] Ofli, F., Chaudhry, R., Kurillo, G., *et al.*: 'Berkeley mhad: a comprehensive multimodal human action database'. IEEE Workshop on Applications of Computer Vision (WACV), 2013, 2013, pp. 53–60

[28] Ravi, D., Wong, C., Lo, B., *et al.*: 'Deep learning for human activity recognition: a resource efficient implementation on low-power devices'. IEEE 13th Int. Conf. on Wearable and Implantable Body Sensor Networks (BSN), 2016, 2016, pp. 71–76

[29] Hendelman, D., Miller, K., Baggett, C., *et al.*: 'Validity of accelerometry for the assessment of moderate intensity physical activity in the field.', *Med. Science Sports Exercise*, 2000, **32**, (9 Suppl), pp. S442–S449

[30] Crouter, S.E., Churilla, J.R., Bassett, D.R.Jr: 'Estimating energy expenditure using accelerometers', *Eur. J. Appl. Physiol.*, 2006, **98**, (6), pp. 601–612

[31] Hees, V.T., Lummel, R.C., Westerterp, K.R.: 'Estimating activity-related energy expenditure under sedentary conditions using a tri-axial seismic accelerometer', *Obesity*, 2009, **17**, (6), pp. 1287–1292

[32] Altini, M., Penders, J., Amft, O.: 'Estimating oxygen uptake during nonsteady-state activities and transitions using wearable sensors', *IEEE J. Biomed. Health Inf.*, 2016, **20**, (2), pp. 469–475

[33] Altini, M., Penders, J., Vullers, R., *et al.*: 'Estimating energy expenditure using body-worn accelerometers: a comparison of methods, sensors number and positioning', *IEEE J. Biomed. Health Inf.*, 2015, **19**, (1), pp. 219–226

[34] Aggarwal, J., Ryoo, M.: 'Human activity analysis: a review', *ACM Comput. Surv.*, 2011, **43**, (3), p. 16

[35] Leo, M., Medioni, G., Trivedi, M., *et al.*: 'Computer vision for assistive technologies', *Comput. Vis. Image Underst.*, 2017, **154**, pp. 1–15

[36] Guo, G., Lai, A.: 'A survey on still image based human action recognition', *Pattern Recognit.*, 2014, **47**, (10), pp. 3343–3361

[37] Laptev, I.: 'On space-time interest points', *Int. J. Comput. Vis.*, 2005, **64**, (2-3), pp. 107–123

[38] Jia, Y., Shelhamer, E., Donahue, J., *et al.*: 'Caffe: convolutional architecture for fast feature embedding'. Proc. of the ACM Int. Conf. on Multimedia, 2014, pp. 675–678

[39] Oreifej, O., Liu, Z.: 'Hon4d: histogram of oriented 4D normals for activity recognition from depth sequences', In the Proceedings of Comput. Vis. Pattern Recognit., 2013, pp. 716–723

[40] Tao, L., Paiement, A., Damen, D., *et al.*: 'A comparative study of pose representation and dynamics modelling for online motion quality assessment', *Comput. Vis. Image Underst.*, 2016, **148**, pp. 136–152

[41] Laptev, I., Marszałek, M., Schmid, C., *et al.*: 'Learning realistic human actions from movies', In the Proceedings of Comput. Vis. Pattern Recognit., 2008, pp. 1–8

[42] Perronnin, F., Sánchez, J., Mensink, T.: 'Improving the fisher kernel for large-scale image classification'. Eur. Conf. Comput. Vis., 2010, pp. 143–156

[43] Ryoo, M., Rothrock, B., Matthies, L.: 'Pooled motion features for first-person videos', In the Proceedings of Comput. Vis. Pattern Recognit., 2015, pp. 896–904

[44] Dobhal, T., Shitole, V., Thomas, G., *et al.*: 'Human activity recognition using binary motion image and deep learning', *Procedia Comput. Sci.*, 2015, **58**, pp. 178–185

[45] Presti, L.L., La Cascia, M.: '3d skeleton-based human action classification: a survey', *Pattern Recognit.*, 2016, **53**, pp. 130–147

[46] Snoek, C.G., Worring, M., Smeulders, A.W.: 'Early versus late fusion in semantic video analysis'. Proc. of the 13th Annual ACM Int. Conf. on Multimedia, 2005, pp. 399–402

[47] Liu, K., Chen, C., Jafari, R., *et al.*: 'Fusion of inertial and depth sensor data for robust hand gesture recognition', *IEEE Sens. J.*, 2014, **14**, (6), pp. 1898–1903

[48] Wu, J., Cheng, J.: 'Bayesian co-boosting for multi-modal gesture recognition.', *J. Mach. Learn. Res.*, 2014, **15**, (1), pp. 3013–3036

[49] Chen, C., Jafari, R., Kehtarnavaz, N.: 'A real-time human action recognition system using depth and inertial sensor fusion', *IEEE Sens. J.*, 2016, **16**, (3), pp. 773–781

[50] Breiman, L.: 'Stacked regressions', *Mach. Learn.*, 1996, **24**, (1), pp. 49–64

[51] OpenNI organization, OpenNI User Guide (November 2010). Available at http://www.openni.org/documentation

[52] Tran, D., Sorokin, A.: 'Human activity recognition with metric learning'. Eur. Conf. on Computer Vision, 2008, pp. 548–561

[53] Dalal, N., Triggs, B.: 'Histograms of oriented gradients for human detection', *Comput. Vis. Pattern Recognit.*, 2005, **1**, pp. 886–893

[54] Dietterich, T.: 'Machine learning for sequential data: a review', In the Proceedings of Struct. Syntactic Statist. Pattern Recognit., 2002, pp. 15–30

[55] Bouten, C.V., Koekkoek, K., Verduin, M., *et al.*: 'A triaxial accelerometer and portable data processing unit for the assessment of daily physical activity', *IEEE Trans. Biomed. Eng.*, 1997, **44**, (3), pp. 136–147

[56] Chang, C., Lin, C.: 'Libsvm: a library for support vector machines', *ACM Trans. Intell. Syst. Technol.*, 2011, **2**, (3), p. 27